

Exhibit B

From: "Chappell, Richard"
Sent: Monday, May 05, 2008 9:37 AM
To: "Olsen, Roger"
Subject: Re-posted files

Renamed the "Dataset" files "Subdatabase" files and re-posted them - to be consistent with terminology in the report. I'm working on writing up **Step 6: Compilation of a "Clean" Database.**

I have a conference call this morning which may take awhile.

From: "Chappell, Richard"
Sent: Wednesday, May 07, 2008 10:41 AM
To: "Olsen, Roger"
Subject: Step 10 write-up
Attachments: RWC_Step_10.doc

Roger,

For your review - there was no previous text for this section. I'm now working on Step 11. Also, had to make a minor change to the SD 1 scree, loadings, sizes file - so am re-posting it now.

I have a (brief, hopefully) conference call at 10am then will call you to discuss where we are.

Rick

Step 10: Identify Major Principal Components

The total variability (or variance) in a multivariate dataset is a function of the number of variables and their individual variances. If the variables exhibit no inter-relationships or mutual correlations then the proportion or percentage of the total variance explained by or accounted for by each variable would be the same. For example, the percentage of the total variance accounted for by each variable in a dataset with $i = 26$ variables, given no mutual correlations, would be $(1/i) \times 100 = (1/26) \times 100 = 3.85\%$. However, this is not true for a multivariate dataset where the variables exhibit at least some degree of mutual correlation. Principal components analysis (PCA) is a commonly used multivariate statistical method for identifying these mutual correlations, if present, and re-apportioning the individual variances accordingly.

PCA operates by transforming a dataset with a large number of variables, ostensibly with inherent mutual correlations, to a new set of uncorrelated reference variables called principal components or PCs. The number of PCs is the same as the number of original variables. However, the apportionment of the total variance among the PCs will depend not only on the number of PCs but on the mutual correlations exhibited by the original variables that comprise the PCs. Given mutual correlations, the objectives of PCA are to: (1) identify those PCs that explain or account for relatively high percentages of the total variance in a dataset, and (2) examine these PCs in order to interpret meaningful relationships among the samples in the dataset. These objectives can only be met by PCA in those cases where the variables exhibit mutual correlations – and hence the dimensionality of the variables in a multivariate dataset can be reduced to a smaller number of significant PCs – and where these PCs exhibit relationships among the samples from which meaningful interpretations are possible. The term “significant” in this context means that a relatively high percentage of the total variance is accounted for (explained) by a small number of PCs.

Experience has shown that the objectives of PCA can be met in a dataset or environmental system dominated by a relatively few number of source impacts that exhibit mutual correlations among their variables. In such cases a correspondingly high percentage of the total variance is explained by only a few PCs, typically 2-3 PCs. This is the reason why EDAnalyzer only extracts (for examination) the top or most significant five PCs: if the top five PCs do not account for a high percentage of the total variance in the system then there is little hope of interpreting meaningful relationships among the samples. In PCA, the PCs are sorted according to the percentage of total variance explained, i.e., from those PCs that account for the highest percentage to those that account for the lowest percentage. One then examines these percentages in order to identify the significant PCs, if any. Most commonly, the percentages are examined graphically.

Many different PCA runs were conducted during this investigation, some of which have been classified as “sensitivity” analyses and some of which have been classified as “investigative” analyses. Those classified as sensitivity analyses were designed to evaluate the sensitivity on the PCA of using certain different variable sets or sample groups. The sensitivity analyses and their results are discussed in more detail in Step 13.

The investigative analyses were design for more direct analysis and interpretation relative to identification of source signatures in the watershed. From the investigative PCA runs, four have been selected (two for water samples labeled SW 3 and SW 17, and two for solids samples labeled SD 1 and SD 6) as the most critical to the investigation or project objectives. Hence the results of these four PCA runs are presented in detail in this report. Aside from their importance, these four runs are also representative of the method used to examine the significance of the PCs, as discussed above, and therefore will be used as such in this section.

One method of displaying the significance of the PCs graphically is a point plot of the percent variance explained versus each PC, where the number of PCs is equal to the total number of original variables – and hence one can show how the variances differ from a corresponding alternate case of no mutual correlations. Such a plot is known as a scree plot, the term “scree” meaning “rubble at the bottom of a cliff” and referring to the random noise in the dataset as the number of PCs increases. Figure ____ shows a scree plot for PCA run SW 3, which contained 26 variables and hence corresponds to 26 PCs, PC 1 through PC 26 on the plot. As shown, the top five PCs (PC 1 through PC 5; indicated with blue symbols) each account for more than $(1/i) \times 100 = (1/26) \times 100 = 3.85\%$ of the total variance in the dataset, the amount attributable to random noise, and hence are considered significant. The amount of variance actually explained by the top five PCs for SW 3 is 74.1%, a significant proportion of the total variance and a significant reduction in dimensionality: from 26 variables explaining 100% of the variance to 5 PCs explaining 74.1%. The remaining variance, $(100 - 74.1) \times 100 = 25.9\%$, is considered to be random noise. An alternate way of displaying this same information is a scree plot in the form of a bar graph, as shown in Figure ____ for SW 3. On the bar graph, the percentage of the total variance explained by the top five PCs are each indicated, i.e., 38.9% (PC 1), 18.2% (PC 2), 7.6% (PC 3), 5.3% (PC 4), and 4.9% (PC5), which indicates that PC 1 and PC 2 are relatively the most important of the five. Similar plots for the other PCA runs are provided in Figures ____ through _____. These all clearly show that the top five PCs are significant, and that the top two PCs are the most significant.

For the two water PCA runs (SW 3 and SW 17), there is no particular advantage of one scree plot version over the other: they both show the same information. However, for the two solids PCA runs (SD 1 and SD 6), the bar graph version has the advantage of also showing an alternate PCA rotation (called the Varimax rotation) that proved useful for additional interpretation of the sample PC scores, as discussed further in this report. As shown on the corresponding figures, the varimax rotation apportions the percentage of total variance differently; however, the total variance explained by the top five PCs is the same: in the case of SD 1, 81.4%, and in the case of SD 6, 81.7%. Again, as in all PCA runs, the PCA in all cases successfully reduced the dimensionality of the datasets from a large number of original variables to a relatively few significant PCs, hence allowing for meaningful interpretations of source impacts in the watershed.

→need 8 figures.